How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review

Year: 2022 | Citations: 127 | Authors: Astrid Bertrand, Rafik Belloum, Winston Maxwell, James R. Eagan

Abstract

The field of eXplainable Artificial Intelligence (XAI) aims to bring transparency to complex AI systems. Although it is usually considered an essentially technical field, effort has been made recently to better understand users' human explanation methods and cognitive constraints. Despite these advances, the community lacks a general vision of what and how cognitive biases affect explainability systems. To address this gap, we present a heuristic map which matches human cognitive biases with explainability techniques from the XAI literature, structured around XAI-aided decision-making. We identify four main ways cognitive biases affect or are affected by XAI systems: 1) cognitive biases affect how XAI methods are designed, 2) they can distort how XAI techniques are evaluated in user studies, 3) some cognitive biases can be successfully mitigated by XAI techniques, and, on the contrary, 4) some cognitive biases can be exacerbated by XAI techniques. We construct this heuristic map through the systematic review of 37 papers-drawn from a corpus of 285-that reveal cognitive biases in XAI systems, including the explainability method and the user and task types in which they arise. We use the findings from our review to structure directions for future XAI systems to better align with people's cognitive processes.